- 强化学习的核心在于，接收环境给出的状态计算对应的奖励，在规定的算法下，更新智能体的策略，以最大奖励。实施强化学习控制的优点在于以下三点：1．支持无模型学习，可以在系统模型未知的情况下进行策略的学习；2．泛用性强，只要规划好一个合适的奖励函数，就能让智能体在复数个环境中学习对应的最优控制策略；3．支持连续和离散系统控制，连续时使用确定性策略，而离散时使用随机性策略进行训练和控制。
- Koopman算子理论是一种可以实现非线性系统全局线性化，并通过线性变换预测系统动态的数学理论。在控制理论中的应用主体由嵌入函数和Koopman算子组成：其中，观测函数指的是一组非线性的变换基底，一般维度都比原非线性系统高，它将非线性系统嵌入到线性系统中，线性系统的基底就由观测函数组成。而koopman算子负责推进系统演化，在这里，koopman算子是数据驱动的，也就是通过接触到真实系统数据，拟合koopman算子，直至能够准确预测系统动态的。
- 在深度Koopman算子上，我采用了嵌入网络和控制网络组合，分别对状态和控制分量进行升维的方法，这里额外控制网络的加入可以有效一直非线性控制变量在以往线性估计中会产生的畸变。同时引入K-step损失函数，评估Koopman网络在K步预测时的性能，通过反向传播的方式优化参数。最后，由价值网络接收升维状态并给出预测，组成了Koopman价值估计框架。
- 将深度Koopman算子加入强化学习，得到如图所示的强化学习框架，Actor网络根据系统状态State给出动作Action；深度Koopman算子会将状态提升至高维线性空间，Critic网络同时接受状态和奖励，用TD算法更新自身参数，并对于策略进行打分并交给Actor以优化Actor网络。该算法中，左边虚线框内的智能体，可以通过与系统不断交互来探索并优化自身，也可以通过经验回放，对得到的数据重复利用。
- 为测试算法性能，在经典的 Lorenz环境中进行测试，由于其连续的特征值谱,通常用于评估基于Koopman的方法。在这里，我采用了LQR中常用的运行成本作为奖励函数，从图中可以，PPO和LQR得到的策略都很不稳定，或效率不高；而我们的算法相较于SAC算法，能够在同样的训练中，取得了更高的周期回报，证明深度Koopman算子能够为流行的SAC算法带来更好的性能。